Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 446

Consonance and the Closure Method in Multiple Testing

Joseph P. Romano, Azeem M. Shaikh and Michael Wolf

September 2009

Consonance and the Closure Method in Multiple Testing

Joseph P. Romano
Departments of Statistics and Economics
Stanford University
romano@stanford.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Michael Wolf
Institute for Empirical Research in Economics
University of Zurich
mwolf@iew.uzh.ch

September 2009

Abstract

Consider the problem of testing s hypotheses simultaneously. In order to deal with the multiplicity problem, the classical approach is to restrict attention to procedures that control the familywise error rate (FWE). Typically, it is known how to construct tests of the individual hypotheses, and the problem is how to combine them into a multiple testing procedure that controls the FWE. The closure method of Marcus et al. (1976), in fact, reduces the problem of constructing multiple test procedures which control the FWE to the construction of single tests which control the usual probability of a Type 1 error. The purpose of this paper is to examine the closure method with emphasis on the concepts of coherence and consonance. It was shown by Sonnemann and Finner (1988) that any incoherent procedure can be replaced by a coherent one which is at least as good. The main point of this paper is to show a similar result for dissonant and consonant procedures. We illustrate the idea of how a dissonant procedure can be strictly improved by a consonant procedure in the sense of increasing the probability of detecting a false null hypothesis while maintaining control of the FWE. We then show how consonance can be used in the construction of some optimal maximin procedures.

KEY WORDS: Multiple Testing, Closure Method, Coherence, Consonance, Familywise Error Rate

JEL CLASSIFICATION NOS: C12, C14.

ACKNOWLEDGMENTS: Research of the first author has been supported by NSF Grant DMS-0707085. Research of the second author has been supported by NSF Grant DMS-0820310. Research of third author has been supported by the Swiss National Science Foundation (NCCR FINRISK, Module A3).

1 Introduction

Consider the general problem of simultaneously testing s individual null hypotheses $H_1, \ldots, H_s$. Data X with distribution $P \in \Omega$ is available, where the parameter space Ω may be a parametric, semiparametric, or nonparametric model for P. In this setting, a general hypothesis H can be viewed as a subset ω of Ω. The problem is to test null hypotheses $H_i : P \in \omega_i$ versus alternative hypotheses $H_i' : P \notin \omega_i$ simultaneously for $i = 1, \ldots, s$. Let $I(P)$ denote the indices of the set of true null hypotheses when P is the true probability distribution, i.e., $i \in I(P)$ if and only if $P \in \omega_i$.

Typically, tests for the individual hypotheses are available and the problem is how to combine them into a suitable multiple testing procedure. The easiest approach is to disregard the multiplicity and simply test each hypothesis in the usual way at level α. However, with such a procedure the probability of one or more false rejections generally increases with s and may be much greater than α.

A classical approach to dealing with this problem is to restrict attention to procedures that control the probability of one or more false rejections. This probability is called the *familywise error rate* (FWE). Here, the term "family" refers to the collection of hypotheses $H_1, \ldots, H_s$ that is being considered for joint testing. Control of the FWE at level α requires that

$$\text{FWE}_P \le \alpha \text{ for all } P \in \Omega \ ,$$

where

$$\text{FWE}_P = P\Big\{\text{reject any } H_i \text{ with } i \in I(P)\Big\} \ .$$

Note that we require $\text{FWE}_P \le \alpha$ for all $P \in \Omega$. Control of the familywise error rate in this sense is called *strong* control of the FWE to distinguish it from *weak* control, where $\text{FWE}_P \le \alpha$ is only required to hold when all null hypotheses are true, i.e., when $P \in \bigcap_{1 \le i \le s} \omega_i$. Since weak control is of limited use in multiple testing, control will always mean strong control for the remainder of this paper. A quite broad treatment of methods that control the FWE is presented in Hochberg and Tamhane (1987).

In order to devise a procedure which controls the FWE, the closure method of Marcus et al. (1976) reduces the problem to constructing single tests that control the usual probability of a Type 1 error. Specifically, for a subset $K \subseteq \{1, \ldots, s\}$, define the intersection hypothesis

$$H_K : P \in \omega_K \ , \tag{1}$$

where

$$\omega_K = \bigcap_{i \in K} \omega_i \ .$$

Of course, $H_i = H_{\{i\}}$. Suppose ϕ_K is a level α test of H_K, i.e.,

$$\sup_{P \in \omega_K} E_P[\phi_K] \le \alpha \ . \tag{2}$$

The closure method rejects H_i if and only if H_K is rejected for all subsets K for which $i \in K$. So, in order for H_i to be deemed significant, every intersection hypothesis which includes H_i must be deemed significant. To see why the closure method controls the FWE, we adapt the argument in Theorem 4.1 of Hochberg and Tamhane (1987). Let A be the event that at least one H_i with $i \in I(P)$ is rejected, and let B be the event that $H_{I(P)}$ is rejected. Then A implies B by construction. Hence,

$$\text{FWE}_P = P\{A\} \leq P\{B\} \leq \alpha \ ,$$

which is the desired result. A thorough review of the topic of multiple testing, emphasizing in particular closed tests in a clinical trial setting, can be found in Bauer (1991).

By reducing the problem of controlling the FWE to that of constructing invidividual tests which control the usual probability of a Type 1 error, the closure method therefore gives a very general approach to the construction of multiple test procedures which control the FWE. In fact, Westfall et al. (1999) elevate the closure method to the status of a principle, where Section 8.2 is called "The Closure Principle." This status is further justified by the result of Sonnemann (1982), who shows that all *coherent* multiple testing procedures that control the FWE can be obtained by applying the closure method to some family of tests which control the usual probability of a Type 1 error. Here, by coherent we mean the requirement that the non-rejection of a hypothesis implies the non-rejection of any hypothesis it implies. Sonnemann and Finner (1988) show further that any incoherent procedure can be replaced by a coherent procedure that rejects the same hypotheses and possibly more. Hence, there is no loss in restricting attention to multiple test procedures constructed using the closure method. These results are reviewed in Section 2. Interestingly, one of the findings of the present work is that, even if the individual tests are constructed in some optimal manner, multiple testing decision rules obtained by closure may actually be inadmissible. This finding was previously obtained in Bittman et al. (2009) in a specific context.

Despite the largely *ad hoc* approaches to construction of multiple tests, we wish to develop methods that are both reliable in control of the FWE and which are designed to have good power. Thus, the main problem motivating this work is how to choose tests of H_K when constructing a multiple testing procedure by the closure method. Even in the case $s = 2$, little formal theory exists in the design of tests of H_K. Along the way, the role of the notion of *consonance* becomes pertinent, and the aim is to investigate this concept. A closed testing method is consonant when the rejection of an intersection hypothesis implies the rejection of at least one of its component hypotheses. Here, we mean that H_j is a component of H_i if $\omega_i \subset \omega_j$. For example, a hypothesis specifying $\theta_1 = \theta_2 = 0$ has component hypotheses $\theta_1 = 0$ and $\theta_2 = 0$, and a consonant procedure which rejects $\theta_1 = \theta_2 = 0$ must reject at least one of the two component hypotheses. In Section 3, we show that there is no need to consider dissonant procedures, i.e., procedures which are not consonant, when testing *elementary* hypotheses, as defined later. Indeed, in such a setting, any dissonant procedure can be replaced by a consonant procedure that rejects the same hypotheses and possibly more. We illustrate the idea of how a dissonant procedure can in fact be strictly improved by a consonant procedure in the sense of increasing the probability of detecting a false null hypothesis while maintaining control of the FWE. Finally, in Section 4, we show how consonance can be used in the construction of some optimal maximin procedures.

2 Coherence

We first provide a lemma, which is a converse of sorts to the closure method. Indeed, the closure method starts with a family of tests of H_K to produce a multiple decision rule. Conversely, given any multiple testing decision rule (not necessarily obtained by the closure method), one can use it to obtain tests of H_K for any intersection hypothesis.

Lemma 2.1 *For testing $H_1, \ldots, H_s$, suppose a given multiple testing decision rule controls the FWE at level α. For testing the intersection hypothesis H_K defined by (1), if H_K is equal to H_i for some i, then test H_K by the test of that H_i. Otherwise, if H_K is not a member of the original family $H_1, \ldots, H_s$, consider the following test ϕ_K of H_K defined by: reject H_K if the given multiple testing decision rule rejects any H_i with $i \in K$. Then, ϕ_K controls the usual probability of a Type 1 error at level α for testing H_K, i.e., it satisfies (2).*

PROOF: If H_K is a member of the original family, the result is trivial. Otherwise, suppose H_K is true, i.e., all H_i with $i \in K$ are true. Then, by construction, the probability that H_K is rejected is the probability any H_i with $i \in K$ is rejected using the given decision rule. Since the given decision rule is assumed to control the FWE at level α, the last probability is no bigger than α. ■

Define a family of hypotheses $H_1, \ldots, H_s$ to be *closed* if each intersection hypothesis H_K is a member of the family. The closure of a given family is the family of all intersection hypotheses induced by the given family. In some cases, there is really is nothing to lose by assuming the given family is closed. One reason is that, when applying the closure method, one gets a rule that controls the FWE, not just for the original family, but for the larger closed family. On the other hand, if one is concerned only with the original family of hypotheses, then we shall see that the notion of consonance may play a role in determining the tests of the additional intersection hypotheses.

The closure method always guarantees that the resulting decision rule is *coherent*, i.e., if H_I implies H_K in the sense that $\omega_I \subset \omega_K$, and H_I is not rejected, then H_K is not rejected. So, if H_K is rejected and $\omega_I \subset \omega_K$, then the requirement of coherence means H_I must be rejected. The requirement of coherence is reasonable because if H_K is established as being false, then H_I is then necessarily false as well if $\omega_I \subset \omega_K$. As stated in Hochberg and Tamhane (1987), coherence "avoids the inconsistency of rejecting a hypothesis without also rejecting all hypotheses implying it." Actually, the same concept was already introduced by Lehmann (1957) under the name *compatibility*.

Sonnemann and Finner (1988) show that any incoherent procedure can be replaced by a coherent one which rejects the same hypotheses as the original procedure and possibly more. Hence, there is no need to consider incoherent multiple testing procedures. Together with the result of Sonnemann (1982) showing that all coherent procedures which control the FWE must be obtained by the closure method, we see that there is no loss in restricting attention to procedures obtained by the closure method. Since these results were originally written in German, we state their results below in Theorems 2.1 and 2.2 for completeness and clarity. A nice review can also be found in Finner and Strassburger (2002). Note, however, that we do not assume the family of hypotheses is closed, and this feature will important later when we discuss consonant tests.

Theorem 2.1 *For testing $H_1, \ldots, H_s$, suppose a coherent method controls the FWE at level α. Then, the method is the result of applying the closure method based on some family of tests ϕ_K of H_K satisfying (2). Thus, all coherent decision rules are generated using the closed testing method.*

PROOF: Define tests of an arbitrary intersection hypothesis H_K as in the statement of Lemma 2.1. Applying the closure method to this family of intersection hypotheses, in fact, results in the same decision rule for the original hypotheses. To see this, first note that any hypothesis that is not rejected by the original rule certainly cannot be rejected by the closure method. Moreover, any hypothesis that is rejected by the original given rule is also rejected by the closure method. Indeed, if H_i is rejected by the original method, then H_K must be rejected when $i \in K$. This occurs by construction if H_K is not a member of the original family and by coherence otherwise. ■

Remark 2.1 Note that the requirement of coherence does not restrict the decision rule unless any of the hypotheses imply any of the others. As a simple example, suppose $X = (X_1, \ldots, X_s)$ is multivariate normal with unknown mean vector $(\theta_1, \ldots, \theta_s)$ and known covariance matrix Σ. If H_i specifies $\theta_i = 0$, then no ω_i is contained in any other ω_j. Hence, in this case, the preceding theorem implies that *all* multiple testing decision rules which control the FWE can be obtained by the closure method. ■

It follows from Theorem 2.1 that in order to devise a multiple testing decision rule which controls the FWE, one can restrict attention to procedures based on the closure method, at least if one is willing to rule out incoherent procedures. But the following result of Sonnemann and Finner (1988) says there is no reason to consider incoherent procedures. Henceforth, it will be convenient to assume that all tests are nonrandomized. This may be done without loss of generality by including an auxiliary independent uniform random variable in the definition of X.

Theorem 2.2 *For testing $H_1, \ldots, H_s$, suppose a given incoherent multiple testing method controls the FWE at level α. Then, it can be replaced by a coherent method which controls the FWE at level α and it rejects at least as many hypotheses as the given incoherent one.*

Proof: Suppose H_i is rejected based on a rejection region R_i. Define a new procedure so that H_i is rejected based on the rejection region

$$R'_i = \bigcup_{j:\omega_j \supseteq \omega_i} R_j \ . \tag{3}$$

This new procedure is coherent in the sense that if H_j is rejected, then so is any H_i for which $\omega_i \subset \omega_j$. To see this, simply note that (3) implies that $R'_j \subseteq R'_i$ whenever $\omega_i \subset \omega_j$.

The new procedure also controls the FWE provided that the original procedure does. To appreciate why, suppose a false rejection is made by the new procedure, i.e., $x \in R'_i$ for some i with $P \in \omega_i$. Then, it must be the case that $x \in R_j$ for some j such that $\omega_j \supseteq \omega_i$. Since $P \in \omega_i$, it follows that $x \in R_j$ for some j such that $P \in \omega_j$. In other words, the original procedure also made a false rejection. Control of the FWE therefore follows from the assumption that the original procedure controls the FWE. ■

Remark 2.2 In the above result, if we further assume that there exists i and $P \in \omega_i^c$ such that

$$P\Big\{ \bigcup_{j:\omega_j \supset \omega_i} R_j \setminus R_i \Big\} > 0 \ ,$$

then, then the new procedure is strictly better in the sense that for some $P \in \omega_i^c$, the probability of rejecting H_i is strictly greater under the new procedure than under the old one. Put differently, we must require that there exists i and j such that $\omega_i \subset \omega_j$ and $P\{R_j \setminus R_i\} > 0$ for some $P \in \omega_i^c$. ■

3 Consonance

It follows from Theorems 2.1 and 2.2 that we can safely restrict attention to the construction of multiple testing decision rules by the closure method. However, not all methods generated by applying the closure method are *consonant*. Recall that consonant methods satisfy that, if H_K is rejected, then some H_i with $i \in K$ is rejected. Hochberg and Tamhane (1987) write on p. 46:

> Nonconsonance does not imply logical contradictions as noncoherence does. This is because the failure to reject a hypothesis is not usually interpreted as its acceptance. [...] Thus whereas coherence is an essential requirement, consonance is only a desirable property.

However, a nonconsonant or *dissonant* procedure can leave the statistician in a difficult situation when explaining the results of a study. For example, consider a randomized experiment for testing the efficiency of a drug versus a placebo with two primary endpoints: testing for reduction in headaches and testing for reduction in muscle pain. Suppose H_1 postulates the drug is no more effective than the placebo for reduction of headaches and H_2 postulates the drug is no more effective than the placebo for reduction of muscle pain. If the joint intersection hypothesis $H_{\{1,2\}}$ is rejected, but the statistician cannot reject either of the individual hypotheses, then compelling evidence has not been established to promote a particular drug indication. The net result is that neither hypothesis can be rejected, even though one might conclude that the drug has some beneficial effect. In this way, lack of consonance makes interpretation awkward.

More importantly, we will argue, not merely from an interpretive viewpoint, but from a mathematical statistics viewpoint, that dissonance is undesirable in that it results in decreased ability to reject false null hypotheses. For concreteness, let us consider a classic example.

Example 3.1 (Two-sided Normal Means Problem) For $1 \leq i \leq 2$, let X_i be independent with $X_i \sim N(\theta_i, 1)$. The parameter space Ω for $\theta = (\theta_1, \theta_2)$ is the entire real plane. Let $s = 2$, so there are only two hypotheses, and null hypothesis H_i specifies $\theta_i = 0$, while the alternative specifies $\theta_i \neq 0$. To apply the closure method, suppose the test of H_i is the uniformly most powerful unbiased (UMPU) level α test, which rejects H_i if $|X_i| > z_{1-\frac{\alpha}{2}}$. All that remains is to choose a test of the joint intersection hypothesis $H_{\{1,2\}}$. There are two well-known choices.

(i) *The uniformly most powerful (rotationally) invariant test.* Apply the test that rejects $H_{\{1,2\}}$ if and only if (X_1, X_2) falls in the rejection region $R_{1,2}(\alpha)$ given by

$$R_{1,2}(\alpha) = \left\{ (X_1, X_2) : \ X_1^2 + X_2^2 > c_2(1-\alpha) \right\} ,$$

where $c_d(1-\alpha)$ denotes the $1-\alpha$ quantile of the Chi-squared distribution with d degrees of freedom. This test is also maximin and most stringent; see Section 8.6 of Lehmann and Romano (2005).

(ii) *Stepdown test based on maximum.* Reject $H_{\{1,2\}}$ if and only if

$$\max\big(|X_1|, |X_2|\big) > m_2(1-\alpha) , \tag{4}$$

where $m_s(1-\alpha)$ is the $1-\alpha$ quantile of the distribution of $\max\big(|X_1|, \ldots, |X_s|\big)$ when the X_i are i.i.d. $N(0,1)$.

In both cases, the closed testing method begins by testing $H_{\{1,2\}}$. If $H_{\{1,2\}}$ is retained, there are no rejections, but if it is rejected, then H_i is rejected if $|X_i| > z_{1-\frac{\alpha}{2}}$. It is easy to see that

$$z_{1-\frac{\alpha}{2}} < m_2(1-\alpha) < c_2^{1/2}(1-\alpha) . \tag{5}$$

The rejection region for test (i) is the outside of a disc centered at the origin of radius $c_2^{1/2}(1-\alpha)$, while the rejection region for test (ii) is the outside of a square centered at the origin and having side length $2m_2(1-\alpha)$; see Figure 1. We now refer to the multiple testing procedure which uses test (i) above for the intersection test as procedure (i), and analogously to the multiple testing procedure which uses test (ii) above for the intersection test as procedure (ii). Then, procedure (ii) is consonant whereas procedure (i) is not. It is clearly possible in procedure (i) to reject the intersection hypothesis $H_{\{1,2\}}$, but to not reject either H_1 or H_2. For example, if $\alpha = 0.05$, then $c_2^{1/2}(0.95) = 2.448$; if $X_1 = X_2 = 1.83$, then $X_1^2 + X_2^2 = 6.698 = 2.588^2$, so $H_{\{1,2\}}$ is rejected but neither X_i satisfies $|X_i| > 1.96$.

Of course, it does not follow that procedure (ii) is preferred merely because it is consonant. The point we wish to make immediately is that, procedure (i) can be improved if that goal is to make correct decisions about H_1 and H_2. Moreover, this is true even though each of the tests of H_1, H_2 and $H_{\{1,2\}}$ possess a strong optimality property, which we see does *not* translate into any optimality property for the overall multiple testing procedure.

To appreciate why, note that we may remove from procedure (i) points in the rejection region for testing the intersection hypothesis $H_{\{1,2\}}$ that do not allow for rejection of either H_1 or H_2. By doing so, we can instead include other points in the rejection region that satisfy the constraint that the overall rule be consonant, while still maintaining error control. In this example, this requires our test of the intersection hypothesis $H_{\{1,2\}}$ to have a rejection region which lies entirely in

$$\big\{(X_1, X_2): \ \max(|X_1|, |X_2|) > z_{1-\frac{\alpha}{2}}\big\} \ .$$

Any intersection test satisfying this constraint for testing the intersection hypothesis will result in a consonant procedure when applying the closure method.

To see a concrete way to improve upon procedure (i), consider a rejection region $R'_{1,2}(\alpha)$ of $H_{\{1,2\}}$ of the form

$$R'_{1,2}(\alpha) = \big\{(X_1, X_2): \ X_1^2 + X_2^2 > c'_2(1-\alpha) \ , \max\big(|X_1|, |X_2|\big) > z_{1-\frac{\alpha}{2}}\big\} \ , \tag{6}$$

where the critical value $c'_2(1-\alpha)$ is chosen so that

$$P_{0,0}\big\{R'_{1,2}(\alpha)\big\} = \alpha \ .$$

Clearly, $c'_2(1-\alpha) < c_2(1-\alpha)$, and the resulting procedure is consonant. For an illustration, see Figure 2. Moreover,

$$P_{\theta_1,\theta_2}\big\{\text{reject } H_i \text{ using } R_{1,2}(\alpha)\big\} < P_{\theta_1,\theta_2}\big\{\text{reject } H_i \text{ using } R'_{1,2}(\alpha)\big\} \ .$$

In particular, the new consonant procedure has uniformly greater power at detecting a false null hypothesis H_i. In this way, imposing consonance not only makes interpretation easier, but it provides better discriminating ability. ■

In the previous example of testing independent normal means, note that if the original family of hypotheses had been H_1, H_2 and $H_{\{1,2\}}$, then we could not claim that the new intersection test is a uniform improvement over the original test. Such improvements are only possible when we use intersection tests as a device to apply the closure method. For this reason, we consider the case where the family of hypotheses of interest $H_1, \ldots, H_s$ is the set of *elementary* hypotheses among the closed testing family. Following Finner and Strassburger (2002), a hypothesis H_i is said to be *maximal* or *elementary* among a family of hypotheses if there exists no H_j in the family with $\omega_i \subset \omega_j$. So, in Example 3.1, H_1 and H_2 are the elementary hypotheses among the closed family. In this setting, we show that there is no need to consider dissonant procedures when applying the closure method because any dissonant procedure can be replaced by a consonant one which reaches the same decisions about the hypotheses in question. The main idea is that when applying the closure method, one should construct intersection hypotheses to form the closed family in a consonant manner. In other words, the rejection region of an intersection hypotheses should be chosen so that points in the rejection region lead to the rejection of at least one elementary hypothesis.

Theorem 3.1 *Consider testing $H_1, \ldots, H_s$ by a closed testing procedure. Assume all hypotheses are elementary. If the given procedure is dissonant among the closed family, it can be replaced by a consonant procedure which reaches the same decisions for the individual elementary hypotheses as the original procedure. Thus, there is no advantage to consider dissonant procedures. Specifically, let $S = \{1, \ldots, s\}$ and assume for $K \subseteq S$ the intersection hypothesis H_K for the given procedure is specified by a rejection region R_K. Define a new procedure based on the closure method with rejection region specified by*

$$R'_K = \bigcup_{i \in K} \bigcap_{J \subseteq S, i \in J} R_J \ . \tag{7}$$

Then, the new procedure is consonant, reaches the same decisions as the original procedure, and therefore controls the FWE if the original method does.

Proof: We claim that

$$R'_K = \bigcup_{i \in K} \bigcap_{J \subseteq S, i \in J} R'_J. \tag{8}$$

To prove (8), we first show that

$$R'_K \supseteq \bigcup_{i \in K} \bigcap_{J \subseteq S, i \in J} R'_J. \tag{9}$$

To see this, note that by intersecting over just the set K instead of many sets J in the inner intersection operation in the definition (7), one obtains

$$R'_K \subseteq R_K \ . \tag{10}$$

Replacing R_J with R'_J in the definition (7) by (10) establishes (9).

Next, we show that

$$R'_K \subseteq \bigcup_{i \in K} \bigcap_{J \subseteq S, i \in J} R'_J. \tag{11}$$

Suppose $x \in R'_K$. Then, there must exist $i^* \in K$ such that

$$x \in \bigcap_{J \subseteq S, i^* \in J} R_J \ . \tag{12}$$

It suffices to show that $x \in R'_L$ for any $L \subseteq S$ such that $i^* \in L$. But, for any such L, by only taking the union in the definition of R'_L in (7) over just i^* and not all $i \in L$, we have that

$$R'_L \supseteq \bigcap_{J \subseteq S, i^* \in J} R_J \ . \tag{13}$$

But, (12) and (13) immediately imply $x \in R'_L$, as required.

The relationship (8) shows the closure method based on the new procedure is consonant. Indeed, (8) states that R'_K consists exactly of those x for which the closure method based on all the R'_J leads to rejection of some H_i with $i \in K$. Specifically, if $x \in R'_K$, then, for some $i^* \in K$,

$$x \in \bigcap_{J \subseteq S, i^* \in J} R'_J \ ,$$

so that H_{i^*} is rejected by the closure method based on the new procedure.

Finally, we argue that both procedures lead to the same decisions. By (10), the new procedure certainly cannot reject any more hypotheses than the original. So, it suffices to show that if a hypothesis, say H_{i^*} is rejected by the original closed testing procedure when x is observed, that it is also rejected by the new method. But, in order for the original procedure to reject H_{i^*} when x is observed, it must be the case that

$$x \in \bigcap_{J \subseteq S, i^* \in J} R_J \ ,$$

which coupled with (13) shows that $x \in R'_L$ for any $L \subseteq S$ such that $i^* \in L$. The closure method then rejects H_{i^*} for the new procedure as well. ■

Remark 3.1 The preceding theorem only asserts that any dissonant procedure can be replaced with a consonant procedure that leads to the same decisions as the dissonant procedure. In most cases, however, one can strictly improve upon a dissonant procedure, as was done in Example 3.1, by removing points of dissonance from the rejection regions of the intersection hypotheses and adding to these rejection regions points that satisfy the constraint that the overall procedure is consonant. ■

Remark 3.2 It is possible to generalize Theorem 3.1 to situations where the family of hypotheses is any strict subset of the closed family. For example, let $H_i : \theta_i = 0$ for $1 \leq i \leq 3$ and consider testing all null hypotheses in the closed family generated by H_1, H_2 and H_3 except $H_{\{1,2,3\}}$. Theorem 3.1 does not apply in this case since not all hypotheses are elementary. Even so, the idea of consonance can be applied as in the proof of the theorem when choosing how to construct the rejection region for the test of $H_{\{1,2,3\}}$. As before, one should simply choose the rejection region for the test of $H_{\{1,2,3\}}$ so that points in the rejection lead to the rejection region of at least one of the other hypotheses. Any decision rule based on a rejection region for $H_{\{1,2,3\}}$ that does not have this feature can be replaced by one that is at least as good in the sense that it rejects the same hypotheses and possibly more. ■

4 Optimality

We now examine the role of consonance in optimal multiple testing procedures. We begin with the following general result. Note that we do not require the hypotheses of interest to be elementary.

Theorem 4.1 *Consider testing hypotheses $H_1, \ldots, H_s$, with corresponding null hypothesis parameter spaces $\omega_1, \ldots, \omega_s$. Let $S = \{1, \ldots, s\}$. Suppose, for testing the individual intersection hypothesis H_S at level α, the test with rejection region R_S maximizes the minimum power over $\gamma_S \subseteq \omega_S^c$. Also, suppose that when applying the closure method to the multiple testing problem using R_S to test H_S, the overall procedure is consonant. Then, the multiple testing procedure maximizes*

$$\inf_{P \in \gamma_S} P\{\text{reject at least one } H_i\} ,$$

among all multiple testing methods controlling the FWE.

Proof: The desired result follows immediately because consonance implies that the rejection of H_S is equivalent to the rejection of at least one H_i. ■

Thus, the overall procedure inherits a maximin property from a maximin property of the intersection test, as long as the overall procedure is consonant. We illustrate this result with an example.

Example 4.1 (Two-sided Normal Means Problem, continued) For $1 \leq i \leq 2$, let X_i be independent with $X_i \sim N(\theta_i, 1)$. The parameter space Ω for $\theta = (\theta_1, \theta_2)$ is the entire real plane. Let $s = 2$, so there are only two hypotheses, and null hypothesis H_i specifies $\theta_i = 0$, while the alternative specifies $\theta_i \neq 0$. For $\epsilon > 0$, define

$$\gamma_{1,2} = \gamma_{1,2}(\epsilon) = \Big\{(\theta_1, \theta_2) : \text{ at least one } \theta_i \text{ satisfies } |\theta_i| \geq \epsilon\Big\} . \tag{14}$$

Then, for testing $H_{\{1,2\}}$ at level α, it is easy to derive the maximin test against $\gamma_{1,2}(\epsilon)$. To see why, apply Theorem 8.1.1 of Lehmann and Romano (2005) with the least favorable distribution uniform over the four points $(\epsilon, 0)$, $(0, \epsilon)$, $(-\epsilon, 0)$ and $(0, -\epsilon)$. The resulting likelihood ratio test rejects for large values of

$$T = T_\epsilon(X_1, X_2) = \cosh\bigl(\epsilon|X_1|\bigr) + \cosh\bigl(\epsilon|X_2|\bigr) \ , \tag{15}$$

where the hyperbolic cosine function $\cosh(\cdot)$ is given by $\cosh(t) = 0.5 \cdot \bigl(\exp(t) + \exp(-t)\bigr)$. The test has rejection region

$$R_{1,2} = R_{1,2}(\epsilon, \alpha) = \bigl\{(X_1, X_2) : T_\epsilon(X_1, X_2) > c(1-\alpha, \epsilon)\bigr\} \ ,$$

where $c(1-\alpha, \epsilon)$ is the $1-\alpha$ quantile of $T_\epsilon(X_1, X_2)$ under $(\theta_1, \theta_2) = (0, 0)$.

Lemma 4.1 *In the above setup, the test with rejection region $R_{1,2}(\epsilon, \alpha)$ maximizes*

$$\inf_{\theta \in \gamma_{1,2}(\epsilon)} P_{\theta_1,\theta_2}\{\text{reject } H_{\{1,2\}}\}$$

among level α tests of $H_{\{1,2\}}$.

Proof: As is well known, the family of distributions of X_i has monotone likelihood ratio in $|X_i|$, and distribution depending only on $|\theta_i|$. Since T is increasing in each of $|X_i|$, it follows by Lemma 5.1 in the Appendix, with $Y_i = |X_i|$ and $\eta_i = |\theta_i|$, that the power function of this test is an increasing function of $|\theta_i|$, and therefore the power function is minimized over $\gamma_{1,2}(\epsilon)$ at the four points $(\epsilon, 0)$, $(0, \epsilon)$, $(-\epsilon, 0)$ and $(0, -\epsilon)$. By Theorem 8.1.1. of Lehmann and Romano (2005) the uniform distribution over these four points is least favorable and the test is maximin. ■

Next, consider the overall multiple testing procedure based on the closed testing method. Take the above $R_{1,2}(\epsilon, \alpha)$ for the test of $H_{\{1,2\}}$ and for H_i take the usual UMPU test, so that $R_i = \bigl\{(X_1, X_2) : |X_i| > z_{1-\frac{\alpha}{2}}\bigr\}$. Then, if we can show this procedure is consonant, it will maximize

$$\inf_{\theta \in \gamma_{1,2}(\epsilon)} P_{\theta_1,\theta_2}\{\text{reject at least one } H_i\} \ ,$$

among all multiple testing methods controlling the FWE at level α. Moreover, because both H_1 and H_2 are false if $(\theta_1, \theta_2) \in \gamma_{1,2}$, we would further be able to claim the procedure maximizes

$$\inf_{\theta \in \gamma_{1,2}(\epsilon)} P_{\theta_1,\theta_2}\{\text{reject at least one true } H_i\} \ , \tag{16}$$

among all multiple testing methods controlling the FWE at level α.

In fact, a consonant procedure results for some values of ϵ. For large values of ϵ, the test statistic $T_\epsilon(X_1, X_2)$ is approximately equivalent to $\max(|X_1|, |X_2|)$, which does lead to a consonant procedure. Indeed, see Figure 3 for an example with $\epsilon = 3$ and $\alpha = 0.05$. Thus, (16) holds for this consonant procedure when the tests of the individual H_i are based on the rejection regions R_i.

On the other hand, for small values of ϵ rejecting for large values of the statistic $T_\epsilon(X_1, X_2)$ is approximately equivalent to rejecting for large values of $X_1^2 + X_2^2$, which we already showed in Example 3.1 does not lead to an overall consonant test when the tests of the individual H_i are based on the rejection regions R_i. So, we do not expect the theorem to apply for such values of ϵ. See Figure 4 for an example with $\epsilon = 0.25$ and $\alpha = 0.05$. ■

When the construction of the maximin test of $H_{\{1,2\}}$ does not lead to a consonant procedure, as above with $\epsilon = 0.25$, one can still derive an improved consonant procedure over the one using this test for the intersection test. However, we must settle for a slightly more limited notion of optimality. In this case, not only must our procedure satisfy the FWE level constraint, but we additionally restrict attention to procedures based on the closure method where the individual tests of H_i have rejection region $\{(X_1, X_2) : |X_i| > z_{1-\frac{\alpha}{2}}\}$. This constraint of forcing the individual rejection regions to be the UMPU tests does not appear unreasonable, though it is an added assumption. Of course, for the reasons mentioned earlier, the restriction to closed testing methods is no restriction at all. Therefore, rather than finding an overall maximin level α test of $H_{\{1,2\}}$, we must find the maximin level α test of $H_{\{1,2\}}$ subject to the additional consonant constraint that its rejection region satisfies

$$R_{1,2} \subseteq R_1 \cup R_2 \ .$$

Corollary 5.1 in the Appendix, a modest generalization of the usual approach, makes it possible. We illustrate its use with an example.

Example 4.2 (Two-sided Normal Means Problem, continued) For $1 \leq i \leq 2$, let X_i be independent with $X_i \sim N(\theta_i, 1)$. The parameter space Ω for $\theta = (\theta_1, \theta_2)$ is the entire real plane. Let $s = 2$, so there are only two hypotheses, and null hypothesis H_i specifies $\theta_i = 0$, while the alternative specifies $\theta_i \neq 0$. Consider the problem of constructing the maximin test for $H_{\{1,2\}}$ over the region $\gamma_{1,2}(\epsilon)$, defined in (14), subject to the constraint that the rejection region is contained in the region where

$$\max\big(|X_1|, |X_2|\big) > z_{1-\frac{\alpha}{2}} \ .$$

We can apply Corollary 5.1 to determine such a test. As before, the least favorable distribution is uniform over the four points $(\epsilon, 0)$, $(0, \epsilon)$, $(-\epsilon, 0)$, and $(0, -\epsilon)$ and large values of the likelihood ratio is equivalent to large values of the statistic $T_\epsilon(X_1, X_2)$ given in (15). The optimal rejection region for the intersection hypothesis $H_{\{1,2\}}$ is then

$$R'_{1,2}(\epsilon, \alpha) = \big\{T_\epsilon(X_1, X_2) > t(1-\alpha, \epsilon), \ \max\big(|X_1|, |X_2|\big) > z_{1-\frac{\alpha}{2}}\big\} \ ,$$

where the constant $t(1-\alpha, \epsilon)$ is determined so that $P_{0,0}\big\{R'_{1,2}(\epsilon, \alpha)\big\} = \alpha$. ■

5 Appendix

Lemma 5.1 *Suppose $Y_1, \ldots, Y_s$ are mutually independent. Further suppose the family of densities on the real line $p_i(\cdot, \eta_i)$ of Y_i have monotone likelihood ratio in Y_i. Let $\psi = \psi(Y_1, \ldots, Y_s)$ be a nondecreasing function of each of its arguments. Then, $E_{\eta_1,\ldots,\eta_s}\left[\psi(Y_1, \ldots, Y_s)\right]$ is nondecreasing in each η_i.*

Proof: The function $\psi(Y_1, Y_2, \ldots, Y_s)$ is nondecreasing in Y_1 with $Y_2, \ldots, Y_s$ fixed. Therefore, by Lemma 3.4.2 of Lehmann and Romano (2005), $E_{\eta_1}\left[\psi(Y_1, \ldots, Y_s)|Y_2, \ldots, Y_s\right]$ is nondecreasing in η_1. So, if $\eta_1 < \eta_1'$, then

$$E_{\eta_1}\left[\psi(Y_1, \ldots, Y_s)|Y_2, \ldots, Y_s\right] \leq E_{\eta_1'}\left[\psi(Y_1, \ldots, Y_s)|Y_2, \ldots, Y_s\right] .$$

Taking expectations of both sides shows the desired result for η_1. To show the result, whenever $\eta_i \leq \eta_i'$, for $i = 1, \ldots, s$, one can apply the above reasoning successively to each component. ■

We now consider the problem of constructing a maximin test where the test must satisfy the level constraint as well as the added constraint that the rejection region must lie in some fixed set R. In our context, R will be the union of the rejection regions for the individual tests. Denote by ω the null hypothesis parameter space and by ω' the alternative hypothesis parameter space over which it is desired to maximize the minimum power. So, the goal now is to determine the test that maximizes

$$\inf_{\theta\in\omega'} E_\theta\left[\phi(X)\right]$$

subject to

$$\sup_{\theta\in\omega} E_\theta\left[\phi(X)\right] \leq \alpha$$

and to the constraint that the rejection region must lie entirely in a fixed subset R. Let $\{P_\theta : \theta \in \omega \cup \omega'\}$ be a family of probability distributions over a sample space $(\mathcal{X}, \mathcal{A})$ with densities $p_\theta = dP_\theta/d\mu$ with respect to a σ-finite measure μ, and suppose that the densities $p_\theta(x)$ considered as functions of the two variables (x, θ) are measurable $(\mathcal{A} \times \mathcal{B})$ and $(\mathcal{A} \times \mathcal{B}')$, where $\mathcal{B}$ and $\mathcal{B}'$ are given σ-fields over ω and ω'. We have the following result.

Corollary 5.1 *Let Λ, Λ' be probability distributions over $\mathcal{B}$ and $\mathcal{B}'$, respectively. Define*

$$\begin{aligned} h(x) &= \int_\omega p_\theta(x)\, d\Lambda(\theta) \\ h'(x) &= \int_{\omega'} p_\theta(x)\, d\Lambda'(\theta) \ . \end{aligned}$$

Let C and γ be constants such that

$$\varphi_{\Lambda,\Lambda'}(x) = \begin{cases} 1 & \textit{if} \quad h'(x) > Ch(x), \ x \in R \\ \gamma & \textit{if} \quad h'(x) = Ch(x), \ x \in R \\ 0 & \textit{if} \quad h'(x) < Ch(x), \ \text{or } x \in R^c \end{cases}$$

is a size-α test for testing the null hypothesis that the density of X is $h(x)$ versus the alternative that it is $h'(x)$ and such that

$$\Lambda(\omega_0) = \Lambda'(\omega_0') = 1,$$

where

$$\begin{aligned} \omega_0 &= \left\{\theta : \theta \in \omega \text{ and } E_\theta \varphi_{\Lambda,\Lambda'}(X) = \sup_{\theta' \in \omega} E_{\theta'} \varphi_{\Lambda,\Lambda'}(X)\right\} \\ \omega_0' &= \left\{\theta : \theta \in \omega' \text{ and } E_\theta \varphi_{\Lambda,\Lambda'}(X) = \inf_{\theta' \in \omega'} E_{\theta'} \varphi_{\Lambda,\Lambda'}(X)\right\} . \end{aligned}$$

Then, $\varphi_{\Lambda,\Lambda'}$ maximizes $\inf_{\theta\in\omega'} E_\theta \varphi(X)$ among all level-α tests $\phi(\cdot)$ of the hypothesis $H : \theta \in \omega$ which also satisfy $\phi(x) = 0$ if $x \in R^c$, and it is the unique test with this property if it is the unique most powerful level-α test among tests that accept on R^c for testing h against h'.

PROOF: It follows from Lemma 1 in Bittman et al. (2009) that $\varphi_{\Lambda,\Lambda'}$ is the most powerful test for testing h against h', among level α tests ϕ that also satisfy $\phi(x) = 0$ if $x \in R^c$. Let $\beta_{\Lambda,\Lambda'}$ be its power against the alternative h'. The assumptions imply that

$$\sup_{\theta\in\omega} E_\theta \varphi_{\Lambda,\Lambda'}(X) = \int_\omega E_\theta \varphi_{\Lambda,\Lambda'}(X)\, d\Lambda(\theta) = \alpha \ ,$$

and

$$\inf_{\theta\in\omega'} E_\theta \varphi_{\Lambda,\Lambda'}(X) = \int_{\omega'} E_\theta \varphi_{\Lambda,\Lambda'}(X)\, d\Lambda'(\theta) = \beta_{\Lambda,\Lambda'}.$$

Thus, the conditions of Theorem 1 in Bittman et al. (2009) hold, and the result follows. ■

References

Bauer, P. (1991). Multiple testing in clinical trials. *Statistics in Medicine*, 10:871–890.

Bittman, R. M., Romano, J. P., Vallarino, C., and Wolf, M. (2009). Testing multiple hypotheses with common effect. *Biometrika*, 96(2):399–410.

Finner, H. and Strassburger, K. (2002). The partitioning principle: a power tool in multiple decision theory. *Annals of Statistics*, 30:1194–1213.

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. Wiley, New York.

Lehmann, E. L. (1957). A theory of some multiple decision problems (Parts I and II). *Annals of Mathematical Statistics*, 28:547–572.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, third edition.

Marcus, R., Peritz, E., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63:655–660.

Sonnemann, E. (1982). Allgemeine Lösungen multipler Testprobleme. *EDV in Medizin und Biologie*, 13:120–128.

Sonnemann, E. and Finner, H. (1988). Vollständigkeitssätze für multiple Testprobleme. In Bauer, P., Hommel, G., and Sonnemann, E., editors, *Multiple Hypothesenprüfung*, pages 121–135. Springer, Berlin.

Westfall, P. H., Tobias, R. D., Rom, D., and Wolfinger, R. D. (1999). *Multiple Comparisons and Multiple Testing Using SAS*. SAS Institute Inc, Cary, North Carolina.

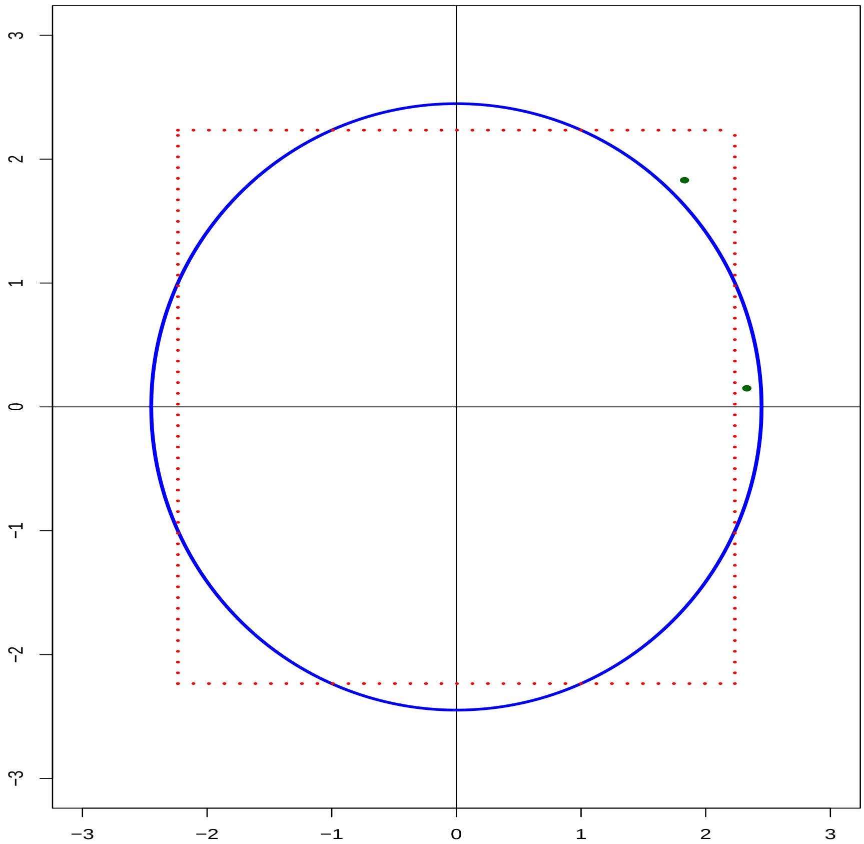

Figure 1: The rejection regions for the two intersection tests of Example 3.1 with nominal level $\alpha = 0.05$. Test (i) rejects for points that fall outside the solid circle with radius 2.448. Test (ii) rejects for points that fall outside the dashed square with length 2×2.234. For example, the point (1.83, 1.83) leads to rejection by test (i) but not by test (ii). On the other hand, the point (2.33, 0.15) leads to the rejection of H_1 by procedure (ii) but not by procedure (i).

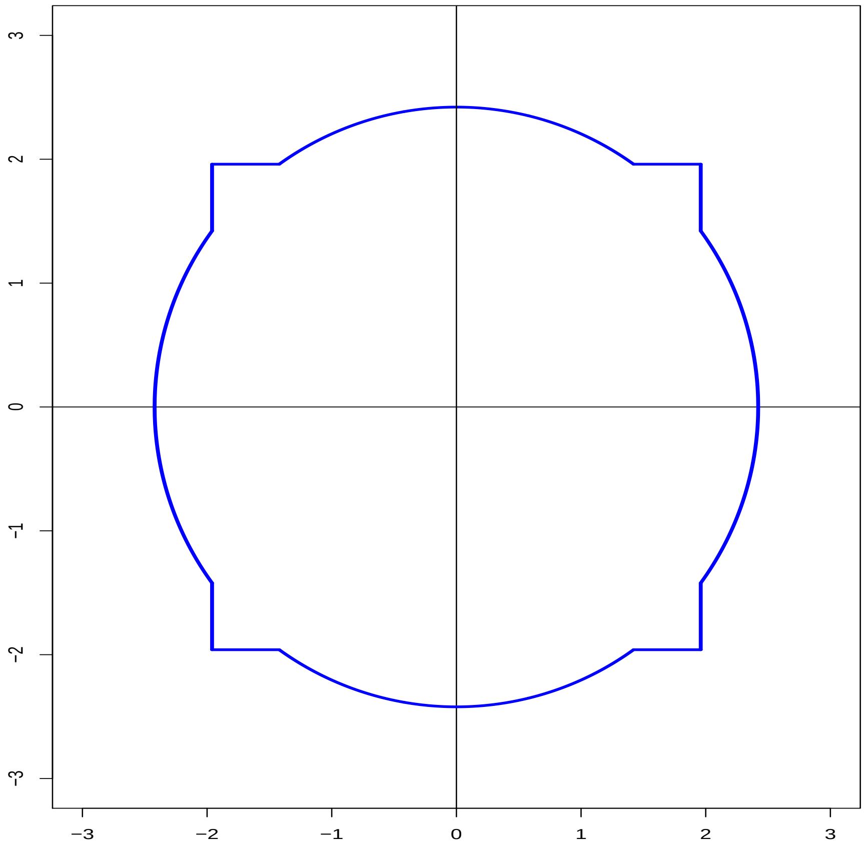

Figure 2: The rejection region $R'_{1,2}(\alpha)$ of the improved procedure (i) of Example 3.1 with nominal level $\alpha = 0.05$; see equation (6). This larger region is obtained as the intersection of the region outside of a circle with radius 2.421 and and the region outside a square with length 2×1.96.

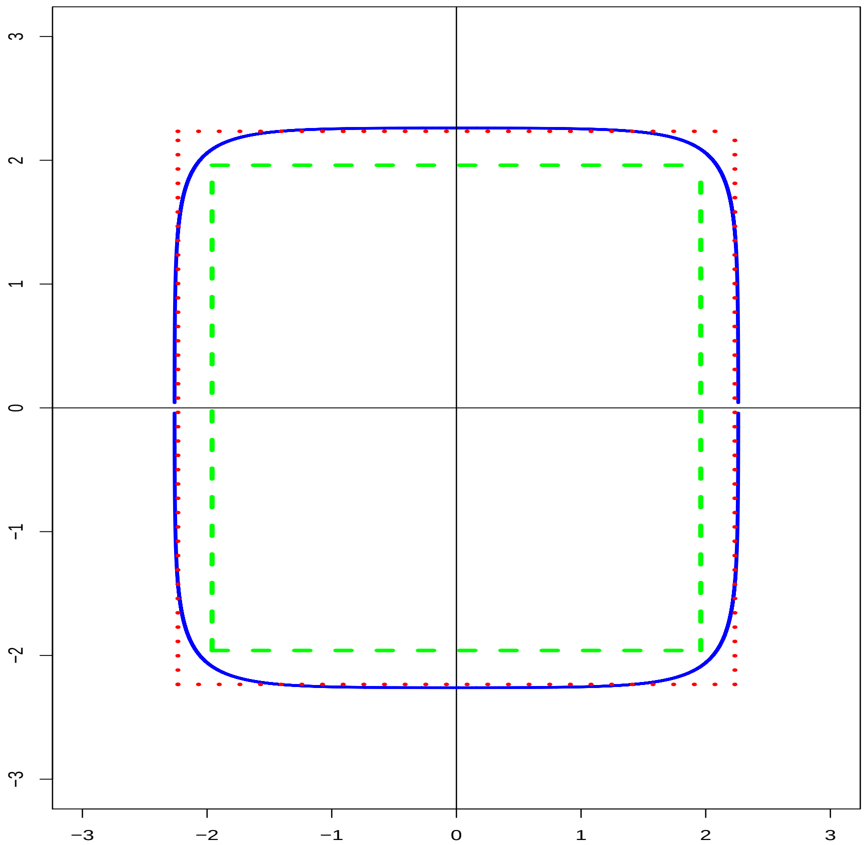

Figure 3: The test of Example 4.1 for $\epsilon = 3$ and nominal level $\alpha = 0.05$. The test rejects for points outside the solid curve. Points outside the inner square with length 2×1.96 lead to rejection of at least one H_i when the individual hypotheses are tested with the usual UMPU test. The outer square with length 2×2.234 is the rejection region of test (ii) of Example 3.1.

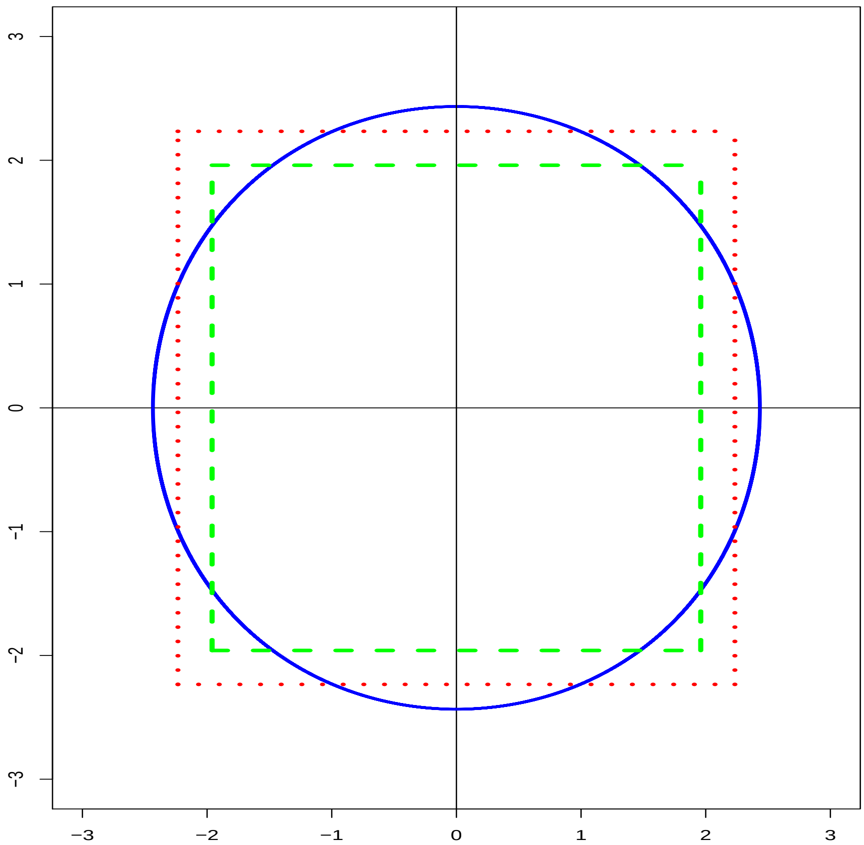

Figure 4: The test of Example 4.1 for $\epsilon = 0.25$ and nominal level $\alpha = 0.05$. The test rejects for points outside the solid curve. Points outside the inner square with length 2×1.96 lead to rejection of at least one H_i when the individual hypotheses are tested with the usual UMPU test. The outer square with length 2×2.234 is the rejection region of test (ii) of Example 3.1.